2020/10

DP

Yurii Nesterov

Inexact high-order proximal-point methods with auxiliary search procedure

CORE
CENTER FOR OPERATIONS RESEARCH AND ECONOMETRICS

CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: immaq-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/core/discussion-papers.html

CORE DISCUSSION PAPER

2020/10

Inexact high-order proximal-point methods with auxiliary search procedure

Yurii Nesterov *

March 4, 2020

Abstract

In this paper, we complement the framework of Bi-Level Unconstrained Minimization (BLUM) [21] by a new pth-order proximal-point method convergent as $O(k^{-(3p+1)/2})$, where k is the iteration counter. As compared with [21], we replace the auxiliary line search by a segment search. This allows us to bound its complexity of by a logarithm of the desired accuracy. Each step in this search needs an approximate computation of a high-order proximal-point operator. Under assumption on boundedness of $(p+1)$st derivative of the objective function, this can be done by one step of the pth-order augmented tensor method. In this way, for $p = 2$, we get a new second-order method with the rate of convergence $O(k^{-7/2})$ and logarithmic complexity of the auxiliary search at each iteration. Another possibility is to compute the proximal-point operator by lower-order minimization methods. As an example, for $p = 3$, we consider the upper-level process convergent as $O(k^{-5})$. Assuming the boundedness of fourth derivative, an appropriate approximation of the proximal-point operator can be computed by a second-order method in logarithmic number of iterations. This combination gives a second-order scheme with much better complexity than the existing theoretical limits.

Keywords: Convex Optimization, tensor methods, proximal-point operator, lower complexity bounds, optimal methods

*Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL). E-mail: Yurii.Nesterov@uclouvain.be. Research results presented in this paper are obtained in the framework of Advanced Grant 788368 of the European Research Council.

1 Introduction

Motivation. In this paper, we develop a new and efficient proximal-point method, which can be used at the upper level in the framework of Bi-Level Unconstrained Minimization (BLUM) [21]. In this approach, we need to decide first on the order of the upper-level procedure, and after that we choose a numerical scheme for an approximate computation of the high-order proximal-point operator. This structure allows a significant flexibility in applying the lower-order methods to the problem classes which were traditionally out of their scope. Consequently, the lower-order methods can demonstrate now much higher rate of convergence breaking down the traditional theoretical limits.

In the first time, the possibility of using the second-order schemes for minimizing functions with Lipschitz-continuous third derivative was demonstrated in [20]. Optimization method in this paper is a second-order implementation of the third-order scheme with the convergence rate $O(k^{-4})$, where k is the iteration counter. Note that the traditional upper bound for the rate of convergence of the second-order methods is $O(k^{-7/2})$ [1, 2, 6]. Of course, the result in [20] does not contradict the theory since it is related only to a subclass of functions, which is usually considered as the natural problem class for the second-order methods (these are functions with Lipschitz-continuous Hessians). However, it destroys the traditional one-to-one correspondence between the problem classes and the orders of applicable minimization schemes. We cannot call anymore a method optimal without mentioning explicitly the corresponding problem class.

In the next paper [21], in order to explain the phenomena of [20], we proposed the framework BLUM, based on the notion of *high-order proximal-point operator.* The proximal approximation of function $f(\cdot)$, defined by

$$\varphi_\lambda(x) \quad = \quad \min_y \left\{ f(y) + \tfrac{1}{2\lambda}\|y-x\|^2 \right\}, \quad \lambda > 0, \tag{1.1}$$

was introduced by Moreau [13]. Later on, Martinet [12] proposed the first proximal-point method based on this operation. The importance of this construction for computational practice was questionable up to the developments of Rockafellar [24], who used the proximal-point iteration in the dual space for justifying the Augmented Lagrangians. This dual scheme was accelerated by Güller [8], who introduced in this method some elements of the Fast Gradient Method from [14]. Some attempts were made by Teboulle and coauthors [25, 9] in studying proximal-point iteration with non-quadratic kernel, concentrating mainly on strongly convex functions. However, during decades this idea was considered as a theoretical tool, which cannot be used in the efficient optimization algorithms.

In [21], it was suggested to regularize the objective function in (1.1) by a power of Euclidean norm $\|\cdot\|^{p+1}$ with $p \geq 1$. We call the corresponding proximal step the *pth-order proximal-point operation.*[1] This terminology is justified by two facts. Firstly, it was shown that the corresponding simple proximal-point method converge as $O\left(k^{-p}\right)$. The rate of convergence of the accelerated version of this method is $O\left(k^{-(p+1)}\right)$. In both cases, we can use appropriate approximations of the proximal point.

Secondly, it appears that this approximation can be computed by one step of the pth order tensor method provided that the pth derivative of the objective function is Lipschitz.

[1] Thus, the standard construction (1.1) corresponds to the first-order proximal points.

Another important contribution of [21] is the definition of the proximal-point operator with *line search*. Using this auxiliary operation, it is possible to achieve the rate of convergence of the order $O(k^{-(3p+1)/2})$. However, in [21] it was assumed that proximal-point line search admits an exact computation. The main goal of this paper is to relax this assumption and justify an implementable version of this method.

Contents. In this paper, we study the methods for solving convex unconstrained optimization problem. It is organized as follows. In Section 2, we introduce a pth-order proximal-point operator with *segment search* ($p \geq 1$) and study its properties. We justify also the rate of convergence $O(k^{-(3p+1)/2})$ of the accelerated method under assumption that the corresponding proximal-point operators admit exact computation. At this stage, we do not need any assumption on the objective function besides convexity. We discuss also the reason why the first-order proximal-point methods do not need auxiliary search.

In Section 3, we define the acceptable approximations for the proximal-point operators and present a conceptual scheme of the accelerated method. We do not specify yet the segment search procedure, assuming only that it satisfies some stopping criterion. The rate of convergence of this scheme is also $O(k^{-(3p+1)/2})$.

In Section 4, we describe the rules of a bisection procedure and show that its length is bounded by logarithmic factors. At each iteration of this scheme, we need to compute approximately a pth-order proximal point.

In the last Section 5, we discuss different strategies for computing approximately the high-order proximal points. In Section 5.1, we show that this can be done by one step of the pth-order augmented tensor method. For that, we need to assume boundedness of $(p+1)$st derivative of the objective function. In this way, we get the second-order methods for functions with bounded third derivative, which have convergence rate $O(k^{-7/2})$.

Finally, in Section 5.2 we present a second-order implementation of the 3rd-order accelerated proximal-point method with segment search. It has the convergence rate $O(k^{-5})$. To the best of our knowledge, this is the fastest rate of convergence known so far for the second-order methods without assumptions on uniform convexity. The number of auxiliary computations of the Hessians at each iteration of this scheme is bounded by logarithmic factors.

Notation and generalities. In what follows, we denote by $\mathbb{E}$ a finite-dimensional real vector space, and by $\mathbb{E}^*$ its dual spaced composed by linear functions on $\mathbb{E}$. For such a function $s \in \mathbb{E}^*$, we denote by $\langle s, x \rangle$ its value at $x \in \mathbb{E}$.

Let us measure distances in $\mathbb{E}$ and $\mathbb{E}^*$ in a Euclidean norm. For that, using a self-adjoint positive-definite operator $B : \mathbb{E} \to \mathbb{E}^*$ (notation $B = B^* \succ 0$), we define

$$\|x\| \;=\; \langle Bx, x \rangle^{1/2}, \quad x \in \mathbb{E}, \quad \|g\|_* \;=\; \langle g, B^{-1} g \rangle^{1/2}, \quad g \in E^*.$$

Sometimes, it will be convenient to treat $x \in \mathbb{E}$ as a linear operator from $\mathbb{R}$ to $\mathbb{E}$, and x^* as a linear operator from $\mathbb{E}^*$ to $\mathbb{R}$. In this case, xx^* is a linear operator from $\mathbb{E}^*$ to $\mathbb{E}$, acting as follows:

$$(xx^*)g \;=\; \langle g, x \rangle x \;\in\; \mathbb{E}, \quad g \in \mathbb{E}^*.$$

For a smooth function $f : \mathbb{E} \to \mathbb{R}$, denote by $\nabla f(x)$ its gradient, and by $\nabla^2 f(x)$ its Hessian evaluated at point $x \in \mathbb{E}$. Note that

$$\nabla f(x) \;\in\; \mathbb{E}^*, \quad \nabla^2 f(x) h \;\in\; \mathbb{E}^*, \quad x, h \in \mathbb{E}.$$

We denote by $\ell_{\bar{x}}(\cdot)$ the linear model of convex function $f(\cdot)$ at point $\bar{x} \in \mathbb{E}$:

$$\ell_{\bar{x}}(x) \;\stackrel{\mathrm{def}}{=}\; f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle \;\leq\; f(x), \quad x \in \mathbb{E}. \tag{1.2}$$

Using the above norm, we can define the standard Euclidean prox-functions

$$d_{p+1}(x) \;=\; \tfrac{1}{p+1}\|x\|^{p+1}, \quad x \in \mathbb{E},$$

where $p \geq 1$ is an integer parameter. These functions have the following derivatives:

$$\begin{aligned} \nabla d_{p+1}(x) \;&=\; \|x\|^{p-1}Bx, \quad x \in \mathbb{E}, \\ \nabla^2 d_{p+1}(x) \;&=\; \|x\|^{p-1}B + (p-1)\|x\|^{p-3}Bxx^*B \\ &\preceq\; p\|x\|^{p-1}B. \end{aligned} \tag{1.3}$$

Note that function $d_{p+1}(\cdot)$ is uniformly convex (see, for example, Lemma 4.2.3 in [17]):

$$d_{p+1}(y) \;\geq\; d_{p+1}(x) + \langle \nabla d_{p+1}(x), y - x \rangle + \tfrac{1}{p+1}\left(\tfrac{1}{2}\right)^{p-1}\|y - x\|^{p+1}, \quad x, y \in \mathbb{E}. \tag{1.4}$$

In this paper, we also work with directional derivatives. For $p \geq 1$, denote by

$$D^p f(x)[h_1, \dots, h_p]$$

the directional derivative of order $p \geq 1$ of function f at x along directions $h_i \in \mathbb{E}$, $i = 1, \dots, p$. Note that $D^p f(x)[\cdot]$ is a *symmetric p-linear form*. Its *norm* is defined in a standard way:

$$\|D^p f(x)\| \;=\; \max_{h_1,\dots,h_p} \left\{ \left| D^p f(x)[h_1, \dots, h_p] \right| : \; \|h_i\| \leq 1, \, i = 1, \dots, p \right\}. \tag{1.5}$$

If all directions $h_1, \dots, h_p$ are the same, we apply notation

$$D^p f(x)[h]^p, \quad h \in \mathbb{E}.$$

Note that, in general, we have (see, for example, Appendix 1 in [22])

$$\|D^p f(x)\| \;=\; \max_{h} \left\{ \left| D^p f(x)[h]^p \right| : \; \|h\| \leq 1 \right\}. \tag{1.6}$$

2 Proximal-point methods with exact search

Consider the following optimization problem:

$$\min_{x \in \mathbb{E}} \; f(x), \tag{2.1}$$

where $f(\cdot)$ is a differentiable convex function. Denote by x^* one of its optimal solutions and let $f^* = f(x^*)$.

All methods presented in this paper are based on the *pth-order proximal-point operators*, defined as follows:

$$\mathrm{prox}^p_{f/H}(\bar{x}) \;=\; \arg\min_{x \in \mathbb{E}} \left\{ f^p_{\bar{x},H}(x) \stackrel{\mathrm{def}}{=} f(x) + H d_{p+1}(x - \bar{x}) \right\}, \tag{2.2}$$

where $H > 0$ and the parameter $p \geq 1$ is called the *order* of this operator.

In [21], it was shown that the rate of convergence of the high-order proximal-point methods can be increased by including in operation (2.2) a *line search*. In this paper, we will see that it is enough to use the proximal-point operators with *segment search*:

$$\mathrm{SProx}^p_{f/H}(\bar{x}, \bar{u}) \;=\; \arg \min_{\substack{x \in \mathbb{E} \\ \tau \in [0,1]}} \Big\{ f(x) + H d_{p+1}(x - \bar{x} - \tau \bar{u}) \Big\} \;\in\; \mathbb{E} \times \mathbb{R}, \tag{2.3}$$

Denote by $x(\bar{x}, \bar{u})$ and $\tau(\bar{x}, \bar{u})$ the solutions of this minimization problem. Since

$$\min_{\substack{x \in \mathbb{E} \\ \tau \in [0,1]}} \Big\{ f(x) + H d_{p+1}(x - \bar{x} - \tau \bar{u}) \Big\} \;=\; \min_{\substack{y \in \mathbb{E} \\ \tau \in [0,1]}} \Big\{ f(y + \tau \bar{u}) + H d_{p+1}(y - \bar{x}) \Big\}, \tag{2.4}$$

the point $x_+ = x(\bar{x}, \bar{u})$ and the value $\tau_+ = \tau(\bar{x}, \bar{u})$ satisfy the first-order optimality condition of the second problem in τ, that is

$$(\tau - \tau_+)\langle \nabla f(x_+), \bar{u} \rangle \;\geq\; 0 \quad \forall \tau \in [0,1]. \tag{2.5}$$

On the other hand, the first-order optimality condition of problem (2.3) in x gives us the following equation:

$$\nabla f(x_+) + H r_+^{p-1} B(x_+ - y_+) \;=\; 0,$$

where $y_+ = \bar{x} + \tau_+ \bar{u}$ and $r_+ = \| x_+ - y_+ \|$. Therefore,

$$\langle \nabla f(x_+), y_+ - x_+ \rangle \;=\; H r_+^{p+1} \;=\; \left[\tfrac{1}{H}\right]^{\frac{1}{p}} \| \nabla f(x_+) \|_*^{\frac{p+1}{p}}. \tag{2.6}$$

Let us assume that we can compute the exact value of operator $\mathrm{SProx}^p_{f/H}(\cdot,\cdot)$. Then, it is possible to solve problem (2.1) by the following efficient optimization scheme.

Exact pth-Order Proximal-Point Method With Segment Search
Initialization. Choose $x_0 \in \mathbb{E}$ and $H > 0$. Define $A_0 = 0$ and $\psi_0(x) = \frac{1}{2}\|x - x_0\|^2$.
Iteration $k \geq 0$. **1.** Compute $v_k = \arg\min_{x \in \mathbb{E}} \psi_k(x)$. Set $u_k = v_k - x_k$. **2.** Compute $(x_{k+1}, \tau_k) = \mathrm{SProx}^p_{f/H}(x_k, u_k)$ and $g_k = \| \nabla f(x_{k+1}) \|_*$. **3.** Compute $a_{k+1} > 0$ from the equation $\frac{a_{k+1}^2}{A_k + a_{k+1}} = \left[\frac{1}{H}\right]^{\frac{1}{p}} g_k^{\frac{1-p}{p}}$. **4.** Set $A_{k+1} = A_k + a_{k+1}$ and update $\psi_{k+1}(x) = \psi_k(x) + a_{k+1} \ell_{x_{k+1}}(x)$.

(2.7)

Using the same arguments as in [21], we can prove that the rate of convergence of this scheme is of the order $O(k^{-(3p+1)/2})$ in terms of the functional residual. In order to see the impact of inexact computation of the proximal step, presented in Section 3, we give here the full proof of the corresponding statement. It is based on two lemmas on positive numbers justified in Appendix (Section 6).

Theorem 1 *Let the sequence $\{x_k\}_{k\geq 0}$ be generated by method (2.7). Then for any $k \geq 1$ we have*

$$f(x_k) - f^* \quad \leq \quad 2^p H R_0^{p+1} \left[1 + \tfrac{2(k-1)}{p+1}\right]^{-(3p+1)/2}, \tag{2.8}$$

where $R_0 = \|x_0 - x^*\|$.

Proof:
Denote $\psi_k^* = \min\limits_{x \in \mathbb{E}} \psi_k(x)$. First of all, let us prove that for all $k \geq 0$, we have

$$A_k f(x_k) + B_k \quad \leq \quad \psi_k^*, \tag{2.9}$$

where $B_k = \frac{1}{2}\left[\frac{1}{H}\right]^{\frac{1}{p}} \sum\limits_{i=0}^{k-1} A_{i+1} g_i^{\frac{p+1}{p}}$. For $k = 0$, we have $A_0 = B_0 = \psi_0^* = 0$. Hence, in this case inequality (2.9) is valid. Assume it is valid for some $k \geq 0$. Then,

$$\begin{aligned}
\psi_{k+1}(x) \quad &= \quad \psi_k(x) + a_{k+1}\ell_{x_{k+1}}(x) \\
&\overset{(2.9)}{\geq} \quad A_k f(x_k) + B_k + \tfrac{1}{2}\|x - v_k\|^2 + a_{k+1}\ell_{x_{k+1}}(x) \\
&\geq \quad A_k f(x_k) + B_k - \tfrac{1}{2}a_{k+1}^2 g_k^2 + a_{k+1}\ell_{x_{k+1}}(v_k).
\end{aligned}$$

Denote $y_k = x_k + \tau_k(v_k - x_k)$ and $\hat{\tau}_k = \frac{a_{k+1}}{A_{k+1}} \in (0,1]$. Then

$$\begin{aligned}
& A_k f(x_k) + a_{k+1}\ell_{x_{k+1}}(v_k) \\
\geq \quad & A_{k+1} f(x_{k+1}) + \langle \nabla f(x_{k+1}), A_k x_k + a_{k+1} v_k - A_{k+1} x_{k+1} \rangle \\
= \quad & A_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), (\hat{\tau}_k - \tau_k) u_k + y_k - x_{k+1} \rangle] \\
\overset{(2.5)}{\geq} \quad & A_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), y_k - x_{k+1} \rangle] \\
\overset{(2.6)}{=} \quad & A_{k+1}\left[f(x_{k+1}) + \left[\tfrac{1}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}}\right].
\end{aligned} \tag{2.10}$$

Thus,

$$\psi_{k+1}^* \quad \geq \quad A_{k+1} f(x_{k+1}) + B_k - \tfrac{1}{2}a_{k+1}^2 g_k^2 + A_{k+1}\left[\tfrac{1}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} \quad = \quad A_{k+1} f(x_{k+1}) + B_{k+1},$$

and the relation (2.9) is proved.

From inequality (2.9) and Step 3) of method (2.7), we get three important consequences. Namely, for all $k \geq 1$, we have

$$
\begin{array}{rrcl}
(A): & f(x_k) - f^* & \leq & \frac{R_0^2}{2A_k}, \\[2ex]
(B): & \left[\frac{1}{H}\right]^{\frac{1}{p}} \sum\limits_{i=1}^{k} A_i g_{i-1}^{\frac{p+1}{p}} & \leq & R_0^2, \\[2ex]
(C): & A_k \left[\frac{1}{H}\right]^{\frac{1}{p}} g_{k-1}^{\frac{1-p}{p}} & = & a_k^2.
\end{array}
\tag{2.11}
$$

Let us obtain from this information a lower bound for the rate of growth of coefficients $\{A_k\}_{k\geq 1}$. This can be done in three steps.

First of all, note that in view of Step 3) of method (2.7), we have

$$
A_{k+1}^{1/2} - A_k^{1/2} \;=\; \frac{a_{k+1}}{A_{k+1}^{1/2}+A_k^{1/2}} \;\geq\; \frac{a_{k+1}}{2A_{k+1}^{1/2}} \;=\; \frac{1}{2}\left[\frac{1}{H}\right]^{\frac{1}{2p}} g_k^{\frac{1-p}{2p}}.
$$

Hence, denoting $\xi_i = g_{i-1}^{\frac{p-1}{2p}}$, $i \geq 1$, we get $A_k \geq \frac{1}{4}\left[\frac{1}{H}\right]^{\frac{1}{p}} \left(\sum\limits_{i=1}^{k} \frac{1}{\xi_i}\right)^2$. On the other hand, in this notation, by inequality $(2.11)_C$, we have

$$
\sum_{i=1}^{k} A_i \xi_i^{\gamma} \;\leq\; H^{\frac{1}{p}} R_0^2,
$$

where $\gamma = \frac{2(p+1)}{p-1}$. Therefore, in view of Lemma 9, we have

$$
\sum_{i=1}^{k} \frac{1}{\xi_i} \;\geq\; \left[\frac{1}{H^{1/p}R_0^2}\right]^{1/\gamma} \left(\sum_{i=1}^{k} A_i^{\frac{1}{1+\gamma}}\right)^{\frac{1+\gamma}{\gamma}} \;=\; H^{-\frac{p-1}{2p(p+1)}} R_0^{-\frac{p-1}{p+1}} \left(\sum_{i=1}^{k} A_i^{\frac{p-1}{3p+1}}\right)^{\frac{3p+1}{2(p+1)}}.
$$

Thus, we get

$$
A_k \;\geq\; \frac{1}{4} H^{-\frac{2}{p+1}} R_0^{-\frac{2(p-1)}{p+1}} \left(\sum_{i=1}^{k} A_i^{\frac{p-1}{3p+1}}\right)^{\frac{3p+1}{p+1}}.
$$

Applying now Lemma 10 with

$$
\theta \;=\; \frac{1}{4} H^{-\frac{2}{p+1}} R_0^{-\frac{2(p-1)}{p+1}}, \quad \alpha \;=\; \frac{p-1}{3p+1}, \quad \beta \;=\; \frac{3p+1}{p+1},
$$

we get $A_k \geq \left(\frac{1}{4}\right)^{\frac{p+1}{2}} \frac{1}{H} R_0^{1-p} \left[\frac{2k+p-1}{p+1}\right]^{\frac{3p+1}{2}}$. It remains to use inequality $(2.11)_A$. $\square$

Remark 1 *In the proof of Theorem 1, it is easy to find the place where we need the segment search. This is the last inequality in the chain (2.10), where we apply the lower bound*

$$
\langle \nabla f(x_{k+1}), (\hat{\tau}_k - \tau_k) u_k \rangle \overset{(2.5)}{\geq} 0.
$$

However, if $p = 1$, then we can achieve this goal in a simpler way. Indeed, in this case the new value of a_{k+1} at Step 3) of the method (2.7) can be obtained before *computing the point x_{k+1}. Consequently, the simple choice $\tau_{k+1} = \hat{\tau}_{k+1}$ ensures validity of the above proof. This possibility is indeed used in Fast Gradient Methods (e.g. [15]), where no line search is needed.*

3 Inexact proximal point computation

Very often, the proximal-point operator (2.2) cannot be computed in a closed form. Instead, we are going to use an approximate solution of this problem obtained by an appropriate optimization procedure. Let us introduce the set of acceptable solutions to problem (2.2), that is

$$\mathcal{A}_H^p(\bar{x},\beta) \;=\; \Big\{ x \in \mathbb{E} : \; \|\nabla f_{\bar{x},H}^p(x)\|_* \leq \beta \|\nabla f(x)\|_* \Big\}, \tag{3.1}$$

where $\beta \in [0,1)$ is a tolerance parameter. Note that $\mathrm{prox}_{f/H}^p(\bar{x}) \in \mathcal{A}_H^p(\bar{x},\beta)$. However, since $\nabla f_{\bar{x},H}^p(\bar{x}) = \nabla f(\bar{x})$, we see that $\bar{x} \notin \mathcal{A}_H^p(\bar{x},\beta)$ unless $\bar{x} = x^*$.

The main properties of the approximate proximal step are given in the following statement (see [21]).

Lemma 1 *Let $T \in \mathcal{A}_H^p(\bar{x},\beta)$. Then*

$$(1-\beta)\|\nabla f(T)\|_* \;\leq\; H\|T-\bar{x}\|^p \;\leq\; (1+\beta)\|\nabla f(T)\|_*, \tag{3.2}$$

$$\langle \nabla f(T), \bar{x} - T\rangle \;\geq\; \tfrac{H}{1+\beta}\|T-\bar{x}\|^{p+1}. \tag{3.3}$$

Moreover, if $\beta \leq \frac{1}{p}$, then

$$\langle \nabla f(T), \bar{x} - T\rangle \;\geq\; \left[\tfrac{1-\beta}{H}\right]^{\frac{1}{p}} \|\nabla f(T)\|_*^{\frac{p+1}{p}}. \tag{3.4}$$

We will need one more property of the approximate proximal step.

Lemma 2 *Let $T \in \mathcal{A}_H^p(\bar{x},\beta)$ with $\beta \in [0,\frac{1}{2})$. Then*

$$\|T-x^*\|^2 \;\leq\; \tfrac{(1-\beta)^2}{1-2\beta}\|\bar{x}-x^*\|^2. \tag{3.5}$$

In particular, if $\beta \leq \frac{3}{8}$, then $\|T-x^*\| \leq \frac{5}{4}\|\bar{x}-x^*\|$.

Proof:
Denote $r = \|T-\bar{x}\|$. Then

$$\begin{aligned}
\|T-x^*\|^2 \;&=\; \|\bar{x}-x^*\|^2 + 2\langle B(T-\bar{x}), T-x^*\rangle - r^2 \\
&\overset{(1.3)}{=}\; \|\bar{x}-x^*\|^2 + \tfrac{2}{Hr^{p-1}}\langle \nabla f_{x,H}^p(T) - \nabla f(T), T-x^*\rangle - r^2 \\
&\overset{(3.1)}{\leq}\; \|\bar{x}-x^*\|^2 + \tfrac{2\beta\|\nabla f(T)\|_*}{Hr^{p-1}}\|T-x^*\| - r^2 \\
&\overset{(3.2)}{\leq}\; \|\bar{x}-x^*\|^2 + \tfrac{2\beta r}{1-\beta}\|T-x^*\| - r^2.
\end{aligned}$$

Maximizing the right-hand side of the last expression in r, we get inequality (3.5). The remaining inequality follows from (3.5) since its right-hand side is monotonically increasing in $\beta \in [0,\frac{1}{2})$. □

Note that the exact computation of the operator $\mathrm{SProx}^p_{f/H}(\cdot,\cdot)$ is also impossible. Therefore, in the remaining part of this section, we analyze an implementable version of method (2.7) based on approximate computations.

Consider the following optimization scheme.

Inexact pth-order Proximal-Point Method With Segment Search

Initialization. Choose $x_0 \in \mathbb{E}$, $\beta \in [0, \frac{3}{3p+2}]$, and $H > 0$. Define $A_0 = 0$ and

$$\psi_0(x) = \tfrac{1}{2}\|x - x_0\|^2.$$

Iteration $k \geq 0$.

1. Compute $v_k = \arg\min\limits_{x\in\mathbb{E}} \psi_k(x)$. Define $u_k = v_k - x_k$.

2. Compute $x_k^0 \in \mathcal{A}_H^p(x_k, \beta)$. **If** $\langle \nabla f(x_k^0), u_k\rangle \geq 0$, **then** define

$$\varphi_k(x) = \ell_{x_k^0}(x), \quad x_{k+1} = x_k^0, \quad g_k = \|\nabla f(x_k^0)\|_*.$$

3. Else, compute $x_k^1 \in \mathcal{A}_H^p(v_k, \beta)$. **If** $\langle \nabla f(x_k^1), u_k\rangle \leq 0$, **then** define

$$\varphi_k(x) = \ell_{x_k^1}(x), \quad x_{k+1} = x_k^1, \quad g_k = \|\nabla f(x_k^1)\|_*. \tag{3.6}$$

4. Else, find the values $0 \leq \tau_k^1 \leq \tau_k^2 \leq 1$ with points $T_k^1 \in \mathcal{A}_H^p(x_k + \tau_k^1 u_k, \beta)$ and $T_k^2 \in \mathcal{A}_H^p(x_k + \tau_k^2 u_k, \beta)$ satisfying the following conditions:

$$\textbf{a)} \quad \beta_k^1 \leq 0 \leq \beta_k^2, \qquad \textbf{b)} \quad \alpha_k(\tau_k^1 - \tau_k^2)\beta_k^1 \leq \tfrac{1}{2}\left[\tfrac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}},$$

where $\beta_k^1 = \langle \nabla f(T_k^1), u_k\rangle$, $\beta_k^2 = \langle \nabla f(T_k^2), u_k\rangle$, $\alpha_k = \frac{\beta_k^2}{\beta_k^2 - \beta_k^1} \in [0,1]$, and

$$g_k = \left[\alpha_k \|\nabla f(T_k^1)\|_*^{\frac{p+1}{p}} + (1-\alpha_k)\|\nabla f(T_k^2)\|_*^{\frac{p+1}{p}}\right]^{\frac{p}{p+1}}.$$

Set $\varphi_k(x) = \alpha_k \ell_{T_k^1}(x) + (1-\alpha_k)\ell_{T_k^2}(x)$ and $x_{k+1} = \alpha_k T_k^1 + (1-\alpha_k)T_k^2$.

5. Compute $a_{k+1} > 0$ from the equation $\frac{a_{k+1}^2}{A_k + a_{k+1}} = \frac{1}{2}\left[\frac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{1-p}{p}}$.

6. Set $A_{k+1} = A_k + a_{k+1}$ and update $\psi_{k+1}(x) = \psi_k(x) + a_{k+1}\varphi_k(x)$.

In this scheme, the Steps 2, 3, and 4 compute an approximate value of the operator $\mathrm{SProx}^p_{f/H}(\cdot,\cdot)$, replacing the Step 2 in method (2.7). Step 4 of method (3.6) is activated only if

$$\langle \nabla f(x_k^0), u_k \rangle \;\; \leq \;\; 0 \; \leq \; \langle \nabla f(x_k^1), u_k \rangle.$$

Thus, the condition **4a)** is satisfied for $\tau_k^1 = 0$ and $\tau_k^2 = 1$. Hence, these values can be used for starting the internal search procedure. Note that the coefficient α_k at this step is chosen to ensure for the vector

$$\nabla \varphi_k \;\; \stackrel{\mathrm{def}}{=} \;\; \alpha_k \nabla f(T_k^1) + (1-\alpha_k) \nabla f(T_k^2)$$

the following equality:

$$\langle \nabla \varphi_k, u_k \rangle \;\; = \;\; \alpha_k \beta_k^1 + (1-\alpha_k)\beta_k^2 \; = \; 0. \tag{3.7}$$

For analyzing the rate of convergence of method (3.6), we need to prove several auxiliary results.

Lemma 3 *For all $k \geq 0$, we have*

$$f(x_k) \;\; \geq \;\; \varphi_k(x_k) \; \geq \; f(x_{k+1}) + \tfrac{1}{2} \left[\tfrac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}}. \tag{3.8}$$

Proof:
The first inequality in (3.8) follows from the convexity of function $f(\cdot)$. In order to prove the second one, we need to consider separately the ways of constructing the function $\varphi_k(\cdot)$ used at Steps 2, 3, and 4.

1. Let $\langle \nabla f(x_k^0), u_k \rangle \geq 0$. Then

$$\varphi_k(x_k) \;\; = \;\; f(x_{k+1}) + \langle \nabla f(x_k^0), x_k - x_k^0 \rangle \;\; \stackrel{(3.4)}{\geq} \;\; f(x_{k+1}) + \left[\tfrac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}}.$$

Thus, the inequality (3.8) is valid.

2. Let $\langle \nabla f(x_k^1), u_k \rangle \leq 0$. Then

$$\begin{aligned} \varphi_k(x_k) \;\; &= \;\; f(x_{k+1}) + \langle \nabla f(x_k^1), x_k - x_k^1 \rangle \; = \; f(x_{k+1}) + \langle \nabla f(x_k^1), v_k - u_k - x_k^1 \rangle \\ &\geq \;\; f(x_{k+1}) + \langle \nabla f(x_k^1), v_k - x_k^1 \rangle \;\; \stackrel{(3.4)}{\geq} \;\; f(x_{k+1}) + \left[\tfrac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}}. \end{aligned}$$

Thus, in this case the inequality (3.8) is also valid.

3. Let us assume now that the function $\varphi_k(\cdot)$ is formed in accordance to the rules of Step 4. Denote $y_k^1 = x_k + \tau_k^1 u_k$, $y_k^2 = x_k + \tau_k^2 u_k$, and $\hat{y}_k = \frac{1}{2}(y_k^1 + y_k^2)$. Then we have

$$\begin{aligned} \varphi_k(x_k) \;\; &= \;\; \alpha_k (f(T_k^1) + \langle \nabla f(T_k^1), x_k - T_k^1 \rangle) + (1-\alpha_k)(f(T_k^2) + \langle \nabla f(T_k^2), x_k - T_k^2 \rangle) \\ &\stackrel{(3.7)}{\geq} \;\; f(x_{k+1}) + \alpha_k \langle \nabla f(T_k^1), \hat{y}_k - T_k^1 \rangle + (1-\alpha_k) \langle \nabla f(T_k^2), \hat{y}_k - T_k^2 \rangle. \end{aligned}$$

Note that

$$\begin{aligned}
\langle \nabla f(T_k^1), \hat{y}_k - T_k^1 \rangle \quad &= \quad \langle \nabla f(T_k^1), y_k^1 - \tfrac{1}{2}(y_k^1 - y_k^2) - T_k^1 \rangle \\
&\overset{(3.4)}{\geq} \quad \left[\tfrac{1-\beta}{H}\right]^{\frac{1}{p}} \|\nabla f(T_k^1)\|_*^{\frac{p+1}{p}} - \tfrac{1}{2}(\tau_k^1 - \tau_k^2)\beta_k^1, \\
\langle \nabla f(T_k^2), \hat{y}_k - T_k^2 \rangle \quad &= \quad \langle \nabla f(T_k^2), y_k^2 + \tfrac{1}{2}(y_k^1 - y_k^2) - T_k^2 \rangle \\
&\overset{(3.4)}{\geq} \quad \left[\tfrac{1-\beta}{H}\right]^{\frac{1}{p}} \|\nabla f(T_k^2)\|_*^{\frac{p+1}{p}} + \tfrac{1}{2}(\tau_k^1 - \tau_k^2)\beta_k^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
f(x_k) \quad &\geq \quad f(x_{k+1}) + \left[\tfrac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} - \tfrac{1}{2}\alpha_k(\tau_k^1 - \tau_k^2)\beta_k^1 + \tfrac{1}{2}(1-\alpha_k)(\tau_k^1 - \tau_k^2)\beta_k^2 \\
&= \quad f(x_{k+1}) + \left[\tfrac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} - \alpha_k(\tau_k^1 - \tau_k^2)\beta_k^1 \overset{(4_c)}{\geq} f(x_{k+1}) + \tfrac{1}{2}\left[\tfrac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}}.
\end{aligned}$$

Thus, inequality (3.8) is proved. □

Corollary 1 *For any $k \geq 0$, we have*

$$\sum_{i=k}^{\infty} g_i^{\frac{p+1}{p}} \quad \leq \quad 2\left[\tfrac{H}{1-\beta}\right]^{\frac{1}{p}} (f(x_k) - f^*). \tag{3.9}$$

Denote $D_0 = \max\limits_{x \in \mathbb{E}}\{\|x - x_0\| : \ f(x) \leq f(x_0)\}$ and $D_* = \max\limits_{x \in \mathbb{E}}\{\|x - x^*\| : \ f(x) \leq f(x_0)\}$. Clearly,

$$D_0 \quad \leq \quad 2D_*. \tag{3.10}$$

In algorithm (3.6), by the way of constructing the estimating functions $\{\psi_k(\cdot)\}_{k \geq 0}$, we have

$$\psi_k(x) \quad \leq \quad A_k f(x) + \tfrac{1}{2}\|x - x_0\|^2, \quad x \in \mathbb{E}, \quad k \geq 0. \tag{3.11}$$

Lemma 4 *Let sequence $\{x_k\}_{k \geq 0}$ generated by method (3.6) be well defined. Then for all $k \geq 0$ we have*

$$A_k f(x_k) + B_k \quad \leq \quad \psi_k^* = \min_{x \in \mathbb{E}} \psi_k(x), \tag{3.12}$$

where $B_k = \frac{1}{4}\left[\frac{1-\beta}{H}\right]^{\frac{1}{p}} \sum\limits_{i=0}^{k-1} A_{i+1} g_i^{\frac{p+1}{p}}$. Moreover, $\|v_k - x_k\| \leq D_0$ and

$$\|x_k - x^*\| \quad \leq \quad D_*, \quad \|v_k - x^*\| \ \leq \ D_*. \tag{3.13}$$

Proof:
From inequality (3.12), we have

$$A_k f(x_k) + B_k + \tfrac{1}{2}\|x - v_k\|^2 \ \leq \ \psi_k(x) \quad \overset{(3.11)}{\leq} \quad A_k f(x) + \tfrac{1}{2}\|x - x_0\|^2, \quad x \in \mathbb{E}. \tag{3.14}$$

Taking in this inequality $x = x_k$, we get $\|v_k - x_k\| \leq \|x_0 - x_k\|$. On the other hand, the choice $x = x_0$ ensures $f(x_k) \leq f(x_0)$. Hence, $\|x_0 - x_k\| \leq D_0$. Finally, the first inequality in (3.13) is valid since $f(x_k) \leq f(x_0)$, and the second one follows from (3.14) with $x = x^*$.

It remains to prove inequality (3.12). Let us do this by induction. For $k = 0$ we have $A_0 = B_0 = \psi_0^* = 0$. Hence, the relation (3.12) is valid. Assume it is valid for some $k \geq 0$. Then

$$\begin{aligned} \psi_{k+1}(x) &= \psi_k(x) + a_{k+1}\varphi_k(x) \;\geq\; A_k f(x_k) + B_k + \tfrac{1}{2}\|x - v_k\|^2 + a_{k+1}\varphi_k(x) \\ &= A_k f(x_k) + B_k + \tfrac{1}{2}\|x - v_k\|^2 + a_{k+1}\varphi_k(x_k) + a_{k+1}\langle \nabla \varphi_k, x - x_k \rangle. \end{aligned}$$

Note that in the case 4a) and 4c), we have $\langle \nabla \varphi_k, u_k \rangle \geq 0$. Therefore,

$$\begin{aligned} & A_k f(x_k) + a_{k+1}\varphi_k(x_k) + a_{k+1}\langle \nabla \varphi_k, x - x_k \rangle \\ \stackrel{(3.8)}{\geq}\; & A_{k+1} f(x_{k+1}) + \tfrac{1}{2} A_{k+1} \left[\tfrac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} + a_{k+1}\langle \nabla \varphi_k, x - x_k \rangle \\ \geq\; & A_{k+1} f(x_{k+1}) + \tfrac{1}{2} A_{k+1} \left[\tfrac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} + a_{k+1}\langle \nabla \varphi_k, x - v_k \rangle. \end{aligned}$$

In the remaining case 4b), we have $\langle \nabla \varphi_k, u_k \rangle \leq 0$. Hence,

$$\begin{aligned} & A_k f(x_k) + a_{k+1}\varphi_k(x_k) + a_{k+1}\langle \nabla \varphi_k, x - x_k \rangle \\ \geq\; & A_{k+1} f(x_{k+1}) + \langle \nabla \varphi_k, A_{k+1}(x_k - x_{k+1}) + a_{k+1}(x - x_k) \rangle \\ =\; & A_{k+1} f(x_{k+1}) + \langle \nabla \varphi_k, A_{k+1}(v_k - x_{k+1}) - A_k(v_k - x_k) + a_{k+1}(x - v_k) \rangle \\ \stackrel{(3.4)}{\geq}\; & A_{k+1} f(x_{k+1}) + A_{k+1} \left[\tfrac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} + a_{k+1}\langle \nabla \varphi_k, x - v_k \rangle. \end{aligned}$$

Thus, in all cases,

$$\begin{aligned} & \psi_{k+1}(x) \\ \geq\; & A_{k+1} f(x_{k+1}) + B_k + \tfrac{1}{2}\|x - v_k\|^2 + \tfrac{1}{2} A_{k+1} \left[\tfrac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} + a_{k+1}\langle \nabla \varphi_k, x - v_k \rangle \\ \geq\; & A_{k+1} f(x_{k+1}) + B_k + \tfrac{1}{2} A_{k+1} \left[\tfrac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} - \tfrac{1}{2} a_{k+1}^2 g_k^2 \\ =\; & A_{k+1} f(x_{k+1}) + B_{k+1} + \tfrac{1}{4} A_{k+1} \left[\tfrac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} - \tfrac{1}{2} a_{k+1}^2 g_k^2 \\ =\; & A_{k+1} f(x_{k+1}) + B_{k+1}. \end{aligned}$$

□

Corollary 2 *Let $R_0 = \|x_0 - x^*\|$. Then*

$$\sum_{i=0}^{\infty} A_{i+1}\, g_i^{\frac{p+1}{p}} \;\leq\; 2 \left[\tfrac{H}{1-\beta}\right]^{\frac{1}{p}} R_0^2. \tag{3.15}$$

Proof:
Indeed, in taking in inequality (3.14) $x = x^*$, we have

$$A_k f(x_k) + B_k + \tfrac{1}{2}\|x^* - v_k\|^2 \quad \leq \quad A_k f^* + \tfrac{1}{2}R_0^2.$$

It remains to use definition of B_k. $\square$

Now we can prove the main statement of this section.

Theorem 2 *Let sequence $\{x_k\}_{k\geq 0}$ generated by method (3.6) be well defined. Then for any $k \geq 1$ we have*

$$f(x_k) - f^* \quad \leq \quad \frac{4^p H R_0^{p+1}}{1-\beta}\left[1 + \frac{2(k-1)}{p+1}\right]^{-\frac{3p+1}{2}}. \tag{3.16}$$

Proof:
In view of inequality (3.12), we have $f(x_k) - f^* \leq \frac{R_0^2}{2A_k}$. Hence, we only need to estimate from below the rate of growth of coefficients $\{A_k\}_{k\geq 0}$.

Using the same arguments as in the proof of Theorem 1, we get

$$A_{k+1}^{1/2} - A_k^{1/2} \quad \geq \quad \frac{a_{k+1}}{2A_{k+1}^{1/2}} \quad = \quad \sqrt{\frac{1}{8}\left[\frac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{1-p}{p}}} \quad \stackrel{\mathrm{def}}{=} \quad c_1 g_k^{\frac{1-p}{2p}}.$$

Denoting $\xi_k = g_k^{\frac{p-1}{2p}}$, we have $\frac{1}{c_1}A_{t+1}^{1/2} \geq \sum\limits_{k=0}^{t} \frac{1}{\xi_k}$ for any $t \geq 0$. On the other hand,

$$B_0 \quad \stackrel{\mathrm{def}}{=} \quad 2\left[\frac{H}{1-\beta}\right]^{\frac{1}{p}} R_0^2 \quad \stackrel{(3.15)}{\geq} \quad \sum_{k=0}^{t} A_{k+1} g_k^{\frac{p+1}{p}} \quad = \quad \sum_{k=0}^{t} A_{k+1}\xi_k^{\frac{2(p+1)}{p-1}}.$$

Therefore, in view of Lemma 9, for $\gamma = \frac{2(p+1)}{p-1}$ we have

$$\frac{1}{c_1}A_{t+1}^{1/2} \quad \geq \quad \frac{1}{B_0^{1/\gamma}}\left(\sum_{k=0}^{t} A_{k+1}^{\frac{1}{1+\gamma}}\right)^{\frac{1+\gamma}{\gamma}} \quad = \quad \frac{1}{B_0^{1/\gamma}}\left(\sum_{k=1}^{t+1} A_k^{\frac{p-1}{3p+1}}\right)^{\frac{3p+1}{2(p+1)}}.$$

Thus, we get the following inequality:

$$A_t \quad \geq \quad \theta\left(\sum_{k=1}^{t} A_k^{\frac{p-1}{3p+1}}\right)^{\frac{3p+1}{p+1}}, \quad t \geq 1, \tag{3.17}$$

where

$$\theta \quad = \quad \frac{c_1^2}{B_0^{2/\gamma}} \quad = \quad \left(\tfrac{1}{2}\right)^{\frac{4p+2}{p+1}}\left[\frac{1-\beta}{H}\right]^{\frac{2}{p+1}} R_0^{-\frac{2(p-1)}{p+1}}.$$

Now we can use Lemma 10 with $\alpha = \frac{p-1}{3p+1}$ and $\beta = \frac{3p+1}{p+1}$. Then $1 - \alpha\beta = \frac{2}{p+1}$, and by inequality (6.4), we get

$$A_k \quad \geq \quad \left(\tfrac{1}{2}\right)^{2p+1}\left[\frac{1-\beta}{H}\right] R_0^{-(p-1)}\left[\frac{p-1}{p+1} + \frac{2}{p+1}k\right]^{\frac{3p+1}{2}}. \qquad \square \tag{3.18}$$

To conclude this section, let us look at the rate of convergence of method (3.6) in terms of the norm of the gradient. Note that the values $\{g_k\}_{k\geq 0}$ generated by method (3.6) are the upper bounds for the norms of the objective function at certain points. Indeed, for Step 2 and Step 3 we have $g_k = \|\nabla f(x_{k+1})\|_*$, and for Step 4, it holds

$$\begin{aligned} g_k &= \left[\alpha_k \|\nabla f(T_k^1)\|_*^{\frac{p+1}{p}} + (1-\alpha_k)\|\nabla f(T_k^2)\|_*^{\frac{p+1}{p}}\right]^{\frac{p}{p+1}} \\ &\geq \min\left\{\|\nabla f(T_k^1)\|_*, \|\nabla f(T_k^2)\|_*\right\}. \end{aligned}$$

Theorem 3 *Let sequence $\{x_k\}_{k\geq 0}$ be generated by method (3.6). The for any $t = 2k$ with $k \geq 1$ we have*

$$g_t^* \stackrel{\mathrm{def}}{=} \min_{0\leq i\leq t-1} g_i \leq 4^p(p+1)^{\frac{p(3p+1)}{2(p+1)}} \cdot \frac{HR_0^p}{1-\beta} \cdot \left[\frac{1}{t}\right]^{\frac{3}{2}p}. \tag{3.19}$$

Proof:
Indeed, in view of inequality (3.16), we have

$$f(x_{k+1}) - f^* \leq \frac{4^p HR_0^{p+1}}{1-\beta}\left[1 + \frac{2k}{p+1}\right]^{-\frac{3p+1}{2}} \leq \frac{4^p HR_0^{p+1}}{1-\beta}\left[\frac{2k}{p+1}\right]^{-\frac{3p+1}{2}}.$$

On the other hand,

$$k(g_t^*)^{\frac{p+1}{p}} \leq \sum_{i=k+1}^{t} g_i^{\frac{p+1}{p}} \stackrel{(3.9)}{\leq} 2\left[\frac{H}{1-\beta}\right]^{\frac{1}{p}} (f(x_{k+1}) - f^*).$$

Combining these two inequalities, we get (3.19). □

It is interesting that the rate of convergence (3.19) is the maximal rate compatible with the rate of growth (3.18) of coefficients $\{A_k\}_{k\geq 0}$. Indeed, if $A_k = O(k^{(3p+1)/2})$, then

$$a_{k+1} = O(k^{(3p-1)/2}), \quad \frac{a_{k+1}^2}{A_{k+1}} = O(k^{3(p-1)/2}).$$

On the other hand, by (3.19) we can guarantee that

$$g_k^{\frac{1-p}{p}} \approx O(k^{3(p-1)/2}),$$

which is compatible with the rule of Step 5 in (3.6). Hence, a higher rate of decrease of the gradients should result in a higher rate of convergence of method (3.6).

4 Complexity of auxiliary search

In this section, we analyze a simple bisection strategy for implementing the Step 4 of method (3.6). Recall that at the beginning of this auxiliary process we have

$$\begin{aligned} \tau_k^1 &= 0, \;\; \beta_k^1 = \langle \nabla f(x_k^0), u_k\rangle \leq 0, \\ \tau_k^2 &= 1, \;\; \beta_k^2 = \langle \nabla f(x_k^1), u_k\rangle \geq 0. \end{aligned}$$

Since for $p = 1$, no auxiliary search is needed (see Remark 1), in this section we assume that $p \geq 2$. Thus, in method (3.6) we have $\beta \leq \frac{3}{3p+2} \leq \frac{3}{8}$, which allows us to use the second statement of Lemma 2.

The following bisection strategy has a single parameter, the tolerance $\beta \in \left[0, \frac{3}{3p+2}\right]$.

$$
\boxed{
\begin{array}{l}
\textbf{Initialization.} \text{ Set } \tau^1_{k,0} = 0,\ \beta^1_{k,0} = \beta^1_k,\ T^1_{k,0} = x^0_k,\ \tau^2_{k,0} = 1,\ \beta^2_{k,0} = \beta^2_k, \\
\text{and } T^2_{k,0} = x^1_k. \text{ Define } \alpha_{k,0} = \frac{\beta^2_{k,0}}{\beta^2_{k,0} - \beta^1_{k,0}} \in [0,1] \text{ and compute} \\
\qquad g_{k,0} = \left[\alpha_{k,0} \| \nabla f(x^0_k) \|_*^{\frac{p+1}{p}} + (1-\alpha_{k,0}) \| \nabla f(x^1_k) \|_*^{\frac{p+1}{p}} \right]^{\frac{p}{p+1}}. \\
\hline
\textbf{Iteration } i \geq 0. \textbf{ If } \alpha_{k,i}(\tau^1_{k,i} - \tau^2_{k,i})\beta^1_{k,i} \leq \frac{1}{2}\left[\frac{1-\beta}{H}\right]^{\frac{1}{p}} g_{k,i}^{\frac{p+1}{p}} \textbf{, then } \boxed{\text{STOP.}} \textbf{ Else,} \\
\textbf{a)} \text{ Compute } \tau^+_{k,i} = \frac{1}{2}(\tau^1_{k,i} + \tau^2_{k,i}) \text{ and } T^+_{k,i} \in \mathcal{A}^p_H(x_k + \tau^+_{k,i} u_k, \beta). \\
\textbf{b)} \text{ Compute } \beta^+_{k,i} = \langle \nabla f(T^+_{k,i}), u_k \rangle. \text{ If } \beta^+_{k,i} \leq 0, \\
\text{then move the left bound } \left\{ \begin{array}{l} \tau^1_{k,i+1} = \tau^+_{k,i},\ \beta^1_{k,i+1} = \beta^+_{k,i},\ T^1_{k,i+1} = T^+_{k,i}, \\ \tau^2_{k,i+1} = \tau^2_{k,i},\ \beta^2_{k,i+1} = \beta^2_{k,i},\ T^2_{k,i+1} = T^2_{k,i}, \end{array} \right. \\
\text{else move the right bound } \left\{ \begin{array}{l} \tau^1_{k,i+1} = \tau^1_{k,i},\ \beta^1_{k,i+1} = \beta^1_{k,i},\ T^1_{k,i+1} = T^1_{k,i}, \\ \tau^2_{k,i+1} = \tau^+_{k,i},\ \beta^2_{k,i+1} = \beta^+_{k,i},\ T^2_{k,i+1} = T^+_{k,i}, \end{array} \right. \\
\textbf{c)}. \text{ Define } \alpha_{k,i+1} = \frac{\beta^2_{k,i+1}}{\beta^2_{k,i+1} - \beta^1_{k,i+1}} \in [0,1] \text{ and compute} \\
\quad g_{k,i+1} = \left[\alpha_{k,i+1} \| \nabla f(T^1_{k,i+1}) \|_*^{\frac{p+1}{p}} + (1-\alpha_{k,i+1}) \| \nabla f(T^2_{k,i+1}) \|_*^{\frac{p+1}{p}} \right]^{\frac{p}{p+1}}.
\end{array}
}
\tag{4.1}
$$

In view of the updating rules in (4.1), for all $i \geq 0$, we have the following relations:

$$
\tau^2_{k,i} - \tau^1_{k,i} \;=\; 2^{-i}, \qquad \beta^1_{k,i} \;\leq\; 0 \;\leq\; \beta^2_{k,i}. \tag{4.2}
$$

Let us find an upper bound for the length of this process.

Note that

$$
\begin{array}{rcl}
0 \;\leq\; -\alpha_{k,i}\beta^1_{k,i} & = & \frac{-\beta^1_{k,i}\beta^2_{k,i}}{\beta^2_{k,i} - \beta^1_{k,i}} \;\leq\; \min\{-\beta^1_{k,i}, \beta^2_{k,i}\} \\[2ex]
& \leq & \|u_k\| \cdot \min\left\{ \|\nabla f(T^1_{k,i})\|_*, \|\nabla f(T^2_{k,i})\|_* \right\}.
\end{array}
$$

Hence, in view of Lemma 4, the left-hand side of the stopping criterion in (4.1) does not exceed

$$
D_0 \min\left\{ \|\nabla f(T^1_{k,i})\|_*, \|\nabla f(T^2_{k,i})\|_* \right\} \cdot 2^{-i}.
$$

On the other hand, $g_{k,i} \geq \min\{\|\nabla f(T^1_{k,i})\|_*, \|\nabla f(T^2_{k,i})\|_*\}$. Hence, the upper bound i_k for the length of the process (4.1) can be found from the following inequality:

$$\frac{HD_0^p}{1-\beta} \cdot 2^{p(1-i_k)} \quad \leq \quad \min\Big\{\|\nabla f(T^1_{k,i})\|_*, \|\nabla f(T^2_{k,i})\|_*\Big\}. \tag{4.3}$$

Theorem 4 *If during the process (4.1), we have*

$$\min\Big\{f(T^1_{k,i}), f(T^2_{k,i})\Big\} \quad \geq \quad f^* + \epsilon, \quad i = 0, \ldots, i_k, \tag{4.4}$$

where $\epsilon > 0$ is the accuracy parameter, then

$$i_k \quad \leq \quad \left(2 + \frac{1}{p}\log_2 \frac{5HD_*^{p+1}}{4(1-\beta)\epsilon}\right)_+, \tag{4.5}$$

where $(\tau)_+ = \max\{\tau, 0\}$ for $\tau \in \mathbb{R}$.

Proof:
In view of inequality (3.13), for any $\tau \in [0,1]$, we have $\|x_k + \tau u_k - x^*\| \leq D_*$. Using now the second statement of Lemma 2, we get

$$\|T^1_{k,i} - x^*\| \quad \leq \quad \tfrac{5}{4}D_*, \quad \|, \quad \|T^2_{k,i} - x^*\| \; \leq \; \tfrac{5}{4}D_*.$$

Therefore,

$$\begin{aligned} \epsilon \quad &\overset{(4.4)}{\leq} \quad \min\Big\{f(T^1_{k,i}) - f^*, f(T^2_{k,i}) - f^*\Big\} \\ &\leq \quad \min\Big\{\langle \nabla f(T^1_{k,i}), T^1_{k,i} - x^*\rangle, \langle \nabla f(T^2_{k,i}), T^2_{k,i} - x^*\rangle\Big\} \\ &\leq \quad \tfrac{5}{4}D_* \min\Big\{\|\nabla f(T^1_{k,i})\|_*, \|\nabla f(T^2_{k,i})\|_*\Big\}. \end{aligned}$$

This means that the stopping criterion in the process (4.1) will be definitely activated for i_k satisfying the inequality $\frac{HD_*^p}{1-\beta} \cdot 2^{p(2-i_k)} \leq \frac{4\epsilon}{5D_*}$, which gives us the bound (4.5). □

5 Practical implementations

The results of Theorem 2 and Theorem 4 give us the following upper bound

$$O\left(\left[\frac{HD_*^{p+1}}{\epsilon}\right]^{\frac{2}{3p+1}} \ln \frac{HD_*^{p+1}}{\epsilon}\right), \quad p \geq 2, \tag{5.1}$$

for the number of computations of the inclusion

$$T \quad \in \quad \mathcal{A}_H^p(x, \beta), \quad x \in \mathbb{E}, \tag{5.2}$$

which is sufficient for getting by method (3.6), (4.1) an ϵ-solution of problem (2.1) in terms of the function value. Note that up to this moment we did not assume anything about the objective function except its convexity. However, we need these assumptions for transforming the bound (5.1) into a bound for the number of calls of oracle. We need also to specify a particular technique for solving the inclusion (5.2).

Thus, we naturally come to the framework of Bi-Level Unconstrained Minimization (BLUM), which was introduced and discussed in Section 4 of [21]. In this framework, at the upper level, we choose the order $p \geq 1$ of the method (3.6), ensuring the complexity bound (5.1). And on the lower level, we choose a numerical method for solving (5.2). In this way, we can construct many interesting combinations. In the remaining part of this section, we consider only two of them.

5.1 Accelerated proximal-point methods with tensor steps

This possibility was already considered in Section 3 of [21] for the slower proximal-point methods with the convergence rate $O\left(k^{-(p+1)}\right)$. Here we just apply these results to the faster scheme (3.6), (4.1).

Denote $M_{p+1}(f) = \sup\limits_{x \in \mathbb{E}} \| D^{p+1} f(x) \|$. For the objective function $f(\cdot)$, we assume that

$$M_{p+1}(f) \quad < \quad +\infty. \tag{5.3}$$

Define the standard *Taylor approximation* of $f(\cdot)$ at point $x \in \mathbb{E}$:

$$\Omega_{x,p}(y) \quad = \quad f(x) + \sum_{k=1}^{p} \tfrac{1}{k!} D^k f(x)[y-x]^k.$$

We need also the *augmented Taylor polynomial*

$$\hat{\Omega}_{x,p,M}(y) \quad = \quad \Omega_{x,p}(y) + \tfrac{M}{(p+1)!} \| y - x \|^{p+1}.$$

If $M \geq M_{p+1}(f)$ then this function provides us with a uniform upper bound for the objective. Moreover, if $M \geq pM_{p+1}(f)$, then function $\hat{\Omega}_{x,p,M}(\cdot)$ is convex (see [18]). Therefore, we are able to compute the tensor step

$$T_{p,M}(x) \quad = \quad \arg\min_{y \in \mathbb{E}} \hat{\Omega}_{x,p,M}(y). \tag{5.4}$$

Let us allow some inexactness in the computation of this point. Namely, we assume that it is possible to compute a point T satisfying the following condition:

$$\| \nabla \hat{\Omega}_{x,p,M}(T) \|_* \quad \leq \quad \tfrac{\gamma}{1+\gamma} \| \nabla \Omega_{x,p}(T) \|_*, \tag{5.5}$$

where $\gamma \in [0, \frac{\beta}{1+\beta})$ is the tolerance parameter.

Let us define now

$$M \quad = \quad \tfrac{1+\beta}{\beta(1-\gamma)-\gamma} M_{p+1}(f), \quad H \; = \; \tfrac{1}{p!} M. \tag{5.6}$$

Then in view of Lemma 3 in [21], we have

$$\| \nabla f(T) + H \nabla d_{p+1}(T - x) \|_* \quad \leq \quad \beta \| \nabla f(T) \|_*.$$

In other words, these values of parameters ensure the inclusion $T \in \mathcal{A}_H^p(x, \beta)$.

Thus, a solution to the inclusion (5.2) can be computed in one step of inexact tensor method provided that we are able to compute its appropriate approximation (5.5). Since the minimization problem (5.4) does not require new calls of oracle of the objective function $f(\cdot)$, the bound (5.1) is valid for the number of calls of pth-order oracle. Note that this bound coincides with the complexity bound for the recently developed tensor methods [5, 7, 10]. However, the auxiliary problem to be solved at each iteration of our method is convex, which should give better practical results.

Let us mention two cases when the solution of problem (5.4) is relatively easy.

- $p = 2$. Then the problem (5.4) coincides with auxiliary problem in Cubic Regularization of Newton Method [23, 16]. It can be solved very efficiently by an appropriate matrix factorization technique. In this case, we take in (5.5) $\gamma = 0$. The number of calls of the second-order oracle in our new method is bounded by (5.1) with $p = 2$, that is
$$O\left(\left[\tfrac{M_3(f) D_*^3}{\epsilon} \right]^{\frac{2}{7}} \ln \tfrac{M_3(f) D_*^3}{\epsilon} \right). \tag{5.7}$$
- $p = 3$. Then the problem (5.4) can be solved by an auxiliary Gradient Method based on Bregmann Distances (see [20] for details). In this case, the computation of an approximate solution (5.2) needs $O(\ln \frac{1}{\epsilon})$ computations of the gradient of the Taylor polynomial of degree tree. The number of calls of third-order oracle in the upper-level method is bounded by (5.1) with $p = 3$, that is
$$O\left(\left[\tfrac{M_4(f) D_*^4}{\epsilon} \right]^{\frac{1}{5}} \ln \tfrac{M_4(f) D_*^4}{\epsilon} \right). \tag{5.8}$$
As it was demonstrated in [20], by an appropriate approximation of the third directional derivative by differences of gradients, we can transform this type of methods into the second-order schemes. In this paper, we do not consider such a possibility since the finite differences are often numerically unstable. In any case, it seems that the next Section 5.2 offers a better alternative.

To the best of our knowledge, the efficient technique for solving the problem (5.4) with the order $p \geq 4$ is not found yet.

5.2 Second-order implementation of third-order method

Let us choose for the upper-level process $p = 3$. This means, that the point $T \in \mathcal{A}_H^3(\bar{x}, \beta)$ is computed as an approximate solution to the problem
$$\min_{x \in \mathbb{E}} \left\{ \varphi(x) \stackrel{\text{def}}{=} f(x) + H d_4(x - \bar{x}) \right\}. \tag{5.9}$$
We are going to solve this problem by a *second-order method.*

Let us assume that $M_4(f) < +\infty$. Define
$$H \;=\; 3M_4(f), \quad \mu \;=\; \tfrac{1}{2}, \quad L \;=\; \tfrac{3}{2}, \quad \varkappa \;=\; \tfrac{1}{3}. \tag{5.10}$$
We need also the scaling function $\rho(x) = \frac{1}{2}\langle \nabla^2 f(\bar{x})(x - \bar{x}), x - \bar{x}\rangle + H d_4(x - \bar{x})$. Then, in view of Theorem 3 in [21], function $\varphi(\cdot)$ satisfies the following *relative non-degeneracy condition* [3, 11]:
$$\mu \nabla^2 \rho(x) \;\preceq\; \nabla \varphi(x) \;\preceq\; L \nabla^2 \rho(x), \quad x \in \mathbb{E}. \tag{5.11}$$

Hence, defining the Bregmann distance $\beta(x,y) = \rho(y) - \rho(x) - \langle \nabla \rho(x), y - x \rangle \geq 0$, we get the following inequality:

$$\varphi(y) \;\; \leq \;\; \varphi(x) + \langle \nabla \varphi(x), y - x \rangle + L\beta(x,y), \quad x, y \in \mathbb{E}.$$

Consequently, we come to the following Gradient Method: set $x_0 = \bar{x}$ and for $k \geq 0$ iterate

$$x_{k+1} \;\; = \;\; \arg\min_{x \in \mathbb{E}} \Big\{ \langle \nabla \varphi(x_k), x - x_k \rangle + L\beta(x_k, x) \Big\}. \tag{5.12}$$

Note that in terms of function $f(\cdot)$, this is mostly a *first-order method*. The Hessian of the objective is computed only once in the beginning of the process in order to define the scaling function $\rho(\cdot)$.

In accordance to [11, 19], method (5.12) has linear rate of convergence:

$$\beta(x_{k+1}, \bar{x}^*) \;\; \leq \;\; (1 - \varkappa)\beta(x_k, \bar{x}_*), \quad k \geq 0, \tag{5.13}$$

where $\bar{x}^*$ is the exact solution of problem (5.9). Thus,

$$\begin{array}{rcl} \frac{H}{16}\|x_k - \bar{x}^*\|^4 & \overset{(1.4)}{\leq} & \beta(x_k, \bar{x}^*) \; \leq \; (1-\varkappa)^k \beta(x_0, \bar{x}^*), \quad k \geq 0, \\ \beta(x_0, \bar{x}_*) & = & \frac{1}{2}\langle \nabla^2 f(\bar{x})(\bar{x} - \bar{x}_*), \bar{x} - \bar{x}_* \rangle + \frac{H}{4}\|\bar{x} - \bar{x}_*\|^4. \end{array} \tag{5.14}$$

Using this fact, we need to find an upper bound for the number of iterations of this scheme, which is sufficient for satisfying the stopping criterion

$$\|\nabla \varphi(x_k)\|_* \;\; \leq \;\; \beta \|\nabla f(x_k)\|_*. \tag{5.15}$$

For this, we need several auxiliary statements. Our main concern is the upper bound for the Hessians of functions $f(\cdot)$ and $\varphi(\cdot)$. We start from a simple statement on second derivatives.

Lemma 5 *For any $x, y \in \mathbb{E}$, we have*

$$\nabla^2 f(y) \;\; \preceq \;\; 2\nabla^2 f(x) + M_4(f)\|x - y\|^2 B. \tag{5.16}$$

Proof:
Indeed, for any $x, h \in \mathbb{E}$, we have

$$\begin{array}{rcl} \nabla^2 f(x+h) & \preceq & \nabla^2 f(x) + D^3 f(x)[h] + \frac{1}{2}M_4(f)\|h\|^2 B, \\ 0 \; \preceq \; \nabla^2 f(x-h) & \preceq & \nabla^2 f(x) - D^3 f(x)[h] + \frac{1}{2}M_4(f)\|h\|^2 B, \end{array}$$

and (5.16) follows. □

In our analysis, we need the following characteristic:

$$M_2 \;\; = \;\; \max_{x \in \mathbb{E}} \Big\{ \|\nabla^2 f(x)\| : \; x \in B(x^*, D_*) \Big\}, \tag{5.17}$$

where $B(x, r) = \{ y \in \mathbb{E} : \; \|y - x\| \leq r \}$. It can be used for bounding the norms of the gradients:

$$\|\nabla f(x)\| \;\; \leq \;\; M_2 D_*, \quad x \in B(x^*, D_*). \tag{5.18}$$

It also gives us a bound for the size of the level set of function $\varphi(\cdot)$.

Lemma 6 *Let $\bar{x} \in B(x^*, D_*)$ and $x \in \mathcal{L}_\varphi = \{x \in \mathbb{E} : \varphi(x) \leq \varphi(\bar{x})\}$. Then*

$$\|x - \bar{x}\| \;\leq\; \hat{R} \stackrel{\mathrm{def}}{=} \left[\tfrac{4M_2D_*}{H}\right]^{1/3}. \tag{5.19}$$

Proof:
Indeed, in this case,

$$\begin{aligned} \tfrac{H}{4}\|x-\bar{x}\|^4 \;&\leq\; f(\bar{x}) - f(x) \;\leq\; \langle \nabla f(\bar{x}), \bar{x} - x\rangle \;\leq\; \|\nabla f(\bar{x})\|_* \cdot \|\bar{x} - x\| \\ &\stackrel{(5.18)}{\leq}\; M_2 D_* \|\bar{x} - x\|. \qquad \square \end{aligned}$$

It is also useful for bounding the Hessians of functions $f(\cdot)$ and $\varphi(\cdot)$.

Lemma 7 *Let $\bar{x} \in B(x^*, D_*)$. Then for any $x \in \mathcal{L}_\varphi$ we have*

$$\begin{aligned} \|\nabla^2 f(x)\| \;&\leq\; M_f \stackrel{\mathrm{def}}{=} 2M_2 + M_4(f)\hat{R}^2, \\ \|\nabla^2 \varphi(x)\| \;&\leq\; M_\varphi \stackrel{\mathrm{def}}{=} L_f + 3H\hat{R}^2. \end{aligned} \tag{5.20}$$

Proof:
Indeed, in view of matrix inequality (5.16), we have

$$\nabla^2 f(x) \;\preceq\; 2\nabla^2 f(\bar{x}) + M_4(f)\|x-\bar{x}\|^2 B \;\stackrel{(5.19)}{\leq}\; (2M_2 + M_4(f)\hat{R}^2)B.$$

It remains to note that $\nabla^2 d_4(x - \bar{x}) \stackrel{(1.3)}{\preceq} 3\|x-\bar{x}\|^2 B \preceq 3\hat{R}^2 B$. $\square$

The size of the constant M_2 can be controlled in different ways. First of all, we can assume that $M_2(f) < +\infty$. Then we can take $M_2 = M_2(f)$. Another possibility is to use the relation (5.16). Then, for any $x \in B(x^*, D_*)$, we have

$$\nabla^2 f(x) \;\preceq\; 2\nabla^2 f(x^*) + M_4(f) D_*^2 B.$$

However, in any case this characteristic has a limited influence on the complexity bounds since it stays inside the logarithms.

Finally, we need one more fact on the properties of the high-order proximal-point operator (2.2) with $p \geq 1$.

Lemma 8 *Let the second-order derivative of function $f(\cdot)$ be bounded on some convex set $\mathcal{L}$ by constant M_f. If both points $\bar{x} \in \mathcal{L}_\varphi$ and $T = prox^p_{f/H}(\bar{x})$ belong to $\mathcal{L}$, then*

$$\|\nabla f(T)\|_* \;\geq\; \frac{H\|\nabla f(\bar{x})\|_*^p}{M_f^p + pH\|\nabla f(\bar{x})\|_*^{p-1}}. \tag{5.21}$$

Proof:
Denote $r = \|T - \bar{x}\|$ and $\bar{g} = \|\nabla f(\bar{x})\|_*$. Then

$$\|\nabla f(T)\|_* \;\geq\; \bar{g} - M_f r.$$

On the other hand, the first-order optimality condition for problem (2.2) reads as follows:

$$\nabla f(T) + H \nabla d_{p+1}(T - \bar{x}) \quad = \quad 0.$$

Hence, $\| \nabla f(T) \|_* \stackrel{(1.3)}{=} r^p$, and we get the following lower bound:

$$\| \nabla f(T) \|_* \quad \geq \quad \max \Big\{ \bar{g} - M_f r, H r^p \Big\}.$$

Since univariate function r^p is convex in $r \geq 0$, the minimum of the right-hand side of this inequality in r is above the following minimum:

$$g_* \stackrel{\mathrm{def}}{=} \min_{r \geq 0} \max \Big\{ \bar{g} - M_f r, H \Big[\hat{r}^p + p \hat{r}^{p-1}(r - \hat{r}) \Big] \Big\} \quad = \quad \frac{H \hat{r}^{p-1}(p\bar{g} - (p-1) M_f \hat{r})}{M_f + p H \hat{r}^{p-1}},$$

where $\hat{r}$ is an arbitrary positive number. Taking $\hat{r} = \frac{\bar{g}}{M_f}$, we get $g_* = \frac{H \bar{g}^p}{M_f^p + p H \bar{g}^{p-1}}$. $\square$

Now we can prove our main result.

Theorem 5 *Let $\bar{x} \in B(x^*, D_*)$ and $f(\bar{x}) - f^* \geq \epsilon > 0$. Then the number of steps of method (5.12) sufficient for activating the stopping criterion (5.15) cannot exceed the value of the order*

$$O \left(\ln \tfrac{1}{\epsilon} + \ln D_* + \ln M_4(f) + \ln M_2 \right). \tag{5.22}$$

Proof:
Since $\bar{x} \in B(x^*, D_*)$, from the condition of the theorem we have

$$\epsilon \quad \leq \quad f(\bar{x}) - f^* \; \leq \; \langle \nabla f(\bar{x}), \bar{x} - x^* \rangle \; \leq \; \| \nabla f(\bar{x}) \|_* \, D_*.$$

Therefore, in view of Lemma 8, we have

$$\| \nabla f(\bar{x}_*) \| \quad \stackrel{(5.18)}{\geq} \quad \frac{H \| \nabla f(\bar{x}) \|_*^3}{M_f^3 + 3H(M_2 D_*)^2} \;\; \geq \;\; \frac{H \epsilon^3}{D_*^3 (M_f^3 + 3H(M_2 D_*)^2)}.$$

Denote by $\delta_k = (1 - \varkappa)^k$ the reduction factor in the convergence rate (5.14). Note that

$$\| x_k - \bar{x}^* \| \quad \stackrel{(5.14)}{\leq} \quad \delta_k^{1/4} \left[\tfrac{8}{H} M_2 \hat{R}^2 + 4 \hat{R}^4 \right]^{\frac{1}{4}}.$$

Since $x_k \in \mathcal{L}_\varphi$, we conclude that

$$\| \nabla f(x_k) \|_* \quad \geq \quad \| \nabla f(\bar{x}^*) \|_* - M_f \| x_k - \bar{x}^* \|.$$

On the other hand, $\| \nabla \varphi(x_k) \|_* \leq M_\varphi \| x_k - \bar{x}^* \|$. Therefore, the stopping criterion (5.15) is satisfied if

$$(M_\varphi + \beta M_f) \| x_k - \bar{x}^* \| \quad \leq \quad \beta \| \nabla f(\bar{x}^*) \|_*.$$

A sufficient condition for this is as follows:

$$\delta_k^{1/4} \left[\tfrac{8}{H} M_2 \hat{R}^2 + 4 \hat{R}^4 \right]^{\frac{1}{4}} (M_\varphi + \beta M_f) \quad \leq \quad \frac{\beta H \epsilon^3}{D_*^3 (M_f^3 + 3H(M_2 D_*)^2)}.$$

Thus, the upper bound for the number of iterations in method (5.12) is of the order (5.22). □

Note that the main assumption of Theorem 5 is the inclusion $\bar{x} \in B(x^*, D_*)$. In view of the relations (3.13), it is valid for all points $\bar{x} = x_k + \tau(v_k - x_k)$, $\tau \in [0,1]$ used in the auxiliary search procedure (4.1).

Thus, we have proved that the third-order accelerated proximal-point method (3.6), (4.1) can be implemented by the second-order method (5.12). This combination has the following performance characteristics.

- Rate of convergence of the upper-level method (3.6) (Theorem 2): $O(k^{-5})$.
- Number of computations of Hessians per iteration of method (3.6) (Theorem 4):

$$N_H \;=\; O(\ln \tfrac{1}{\epsilon} + \ln D_* + \ln M_4(f)).$$

- Number of computations of gradients per iteration of this method (Theorem 5):

$$N_G \;=\; O\Big(N_H \times (\ln \tfrac{1}{\epsilon} + \ln D_* + \ln M_4(f) + \ln M_2)\Big).$$

  Note that the total time for computing the gradients is normally dominated by the time spent for the Hessians. Indeed, usually the Hessians are $(\dim \mathbb{E})$-times more expensive than the gradients. At the same time, very often we have

$$\dim \mathbb{E} \;>\; \ln \tfrac{1}{\epsilon} + \ln D_* + \ln M_4(f) + \ln M_2.$$

6 Appendix

In this section, we present two useful statements on positive numbers.

Lemma 9 *Let the positive numbers $\{\xi_i\}_{i=1}^N$ satisfy the following condition:*

$$\sum_{i=1}^N w_i \, \xi_i^\gamma \;\le\; B, \tag{6.1}$$

where $\gamma \ge 1$ and all weights w_i are positive. Then

$$\sum_{i=1}^N \tfrac{1}{\xi_i} \;\ge\; \tfrac{1}{B^{1/\gamma}} \left(\sum_{i=1}^N w_i^{\frac{1}{1+\gamma}} \right)^{\frac{1+\gamma}{\gamma}}. \tag{6.2}$$

Proof:
Indeed, minimizing the left-hand side of inequality (6.2) with the constraint (6.1), we have the following first-order optimality condition:

$$\tfrac{\lambda}{\xi_i^2} \;=\; w_i \, \xi_i^{\gamma-1}, \quad 1 \le i \le N,$$

where $\lambda > 0$ is the Lagrange multiplier for the active constraint. Thus, $\xi_i = \left(\frac{\lambda}{w_i}\right)^{\frac{1}{\gamma+1}}$, and the multiplier λ can be found from the equation

$$\sum_{i=1}^n w_i \left(\tfrac{\lambda}{w_i}\right)^{\frac{\gamma}{\gamma+1}} \;=\; B.$$

This is $\lambda = \left(\frac{B}{S}\right)^{\frac{1+\gamma}{\gamma}}$, where $S = \sum\limits_{i=1}^{N} w_i^{\frac{1}{1+\gamma}}$. Hence, we get (6.2). $\Box$

Lemma 10 *Let the sequence of positive numbers $\{\eta_k\}_{k\geq 0}$ satisfy the following inequality:*

$$\eta_k \quad \geq \quad \theta \cdot \left(\sum_{i=1}^{k} \eta_i^{\alpha} \right)^{\beta}, \quad k \geq 1, \tag{6.3}$$

where $\theta > 0$, $\alpha \geq 0$, $\beta \geq 0$, and $\alpha\beta < 1$. Then

$$\eta_k \quad \geq \quad \theta^{\frac{1}{1-\alpha\beta}} \Big[\alpha\beta + (1-\alpha\beta)k \Big]^{\frac{\beta}{1-\alpha\beta}}, \quad k \geq 1. \tag{6.4}$$

Proof:

Denote $C_k = \left(\sum\limits_{i=1}^{k} \eta_i^{\alpha} \right)^{1-\alpha\beta}$. Then, in view of (6.3), we have

$$C_{k+1}^{\frac{1}{1-\alpha\beta}} - C_k^{\frac{1}{1-\alpha\beta}} \quad = \quad \eta_{k+1}^{\alpha} \overset{(6.3)}{\geq} \theta^{\alpha} \cdot C_{k+1}^{\frac{\alpha\beta}{1-\alpha\beta}}.$$

Note that for $k = 1$ we have $C_1 = \eta_1^{\alpha(1-\alpha\beta)} \overset{(6.3)}{\geq} \theta^{\alpha}$. On the other hand, since $\frac{1}{1-\alpha\beta} > 1$, we have

$$\begin{aligned} \theta^{\alpha} \cdot C_{k+1}^{\frac{\alpha\beta}{1-\alpha\beta}} \quad &\leq \quad C_{k+1}^{\frac{1}{1-\alpha\beta}} - C_k^{\frac{1}{1-\alpha\beta}} \quad \leq \quad \tfrac{1}{1-\alpha\beta} C_{k+1}^{\frac{1}{1-\alpha\beta}-1} (C_{k+1} - C_k) \\ &= \quad \tfrac{1}{1-\alpha\beta} C_{k+1}^{\frac{\alpha\beta}{1-\alpha\beta}} (C_{k+1} - C_k). \end{aligned}$$

Hence, $C_{k+1} - C_k \geq (1-\alpha\beta)\theta^{\alpha}$, and we get inequality (6.4). $\Box$

References

[1] N. Agarwal, E. Hazan. Lower Bounds for Higher-Order Convex Optimization. arXiv:1710.10329v1 [math.OC] (2017).

[2] Arjevani, O. Shamir, and R. Shiff. Oracle Complexity of Second-Order Methods for Smooth Convex Optimization. arXiv:1705.07260 [math.OC] (2017).

[3] H.H. Bauschke, J. Bolte, M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first order methods revisited and applications. *Math. Oper. Res.* **42**, 330–348 (2016).

[4] M. Baes. Estimate sequence methods: extensions and approximations. *Optimization Online* (2009).

[5] S. Bubeck, Q. Jiang, Y.T. Lee, Y. Li, and A. Sidford. Near-optimal methods for highly smooth convex optimization. *arXiv: 1812.08026v2* [math.OC] 23 Jun 2019.

[6] E. G. Birgin, J. L. Gardenghi, J. M. Martinez, S. A. Santos, and Ph. L. Toint. Worstcase evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, **163**, 359-368 (2017).

[7] A. Gasnikov, E. Gorbunov, D. Kovalev, A. Mohhamed, and E. Chernousova. The global rate of convergence for optimal tensor methods in smooth convex optimization. arXiv: 1809.00382, (2018).

[8] O. Güller. New proximal point algorithms for convex minimization. *SIAM Journal on Control and Optimization*, **14**(5), 877-898 (1992)

[9] A.N. Iusem, B.F. Svaiter, and M. Teboulle. Entropy-like proximal methods in Convex Programming. *Math. oper. res.* **19**(4) 790-814 (1994).

[10] B. Jiang, H. Wang, and S. Zang. An optimal high-order tensor method for convex optimization. arXiv: 1812.06557, (2018).

[11] H. Lu, R. Freund, and Yu. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIOPT*, **28**(1), 333-354 (2018).

[12] B. Martinet. Perturbation des methods d'Optimization. Application., R.A.I.R.O. *Numer. Anal.* (1978).

[13] J.J. Moreau. Proximité et Dualité dans un Espace Hilbertien. *Bull. Soc. Math. France* **93** 273-299 (1965).

[14] Yu. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(\frac{1}{k^2})$. *Doklady AN SSSR* (translated as Soviet Math. Docl.), **269**(3), 543-547 (1983).

[15] Yu. Nesterov. "Smooth minimization of non-smooth functions", *Mathematical Programming* (A), **103** (1), 127-152 (2005).

[16] Yu. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, **112**(1) 159-181 (2008)

[17] Yu. Nesterov. Lectures on Convex Optimization. *Springer* (2018).

[18] Yu. Nesterov. Implementable Tensot Methods in Unconstrained Convex Optimization. *Mathematical Programming*, DOI 10.1007/s10107-019-01449-1 (2019)

[19] Yu. Nesterov. Inexact Basic Tensor Methods. *CORE DP* #2019/23, (2019).

[20] Yu. Nesterov. Superfast second-order methods for Unconstraine Convex Optimization. CORE DP, #2020/7, (January 2020)

[21] Yu. Nesterov. Inexact accelerated high-order proximal-point methods. *CORE Discussion Paper* 2020/08 (February 2020)

[22] Yu. Nesterov, A. Nemirovskii. *Interior point polynomial methods in convex programming: Theory and Applications,*, SIAM, Philadelphia, 1994.

[23] Yu. Nesterov, B. Polyak. "Cubic regularization of Newton's method and its global performance". *Mathematical Programming*, **108**(1), 177-205 (2006).

[24] R.T. Rockafellar. Augmented Lagrangians and Applications of the Proximal Point Algorithm in Convex Programming. *Math. Oper. Res.* **1**, 97-116 (1976).

[25] M. Teboulle. Entropic proximal mapping with applications to Nonlinear Programming. *Math. Oper. Res.* **17**, 670-690 (1992)